Analyse Statistique de Graphes
Block Models pour les graphes multiparties - Applications en écologie et en ethnobiologie

Alya Ben Abdallah & Badis Jaouani
Sorbonne Université

1

Introduction

Les graphes sont devenus un outil très important dans différents domaines tels que la biologie, l'écologie ou encore la sociologie. Différents exemples de graphes peuvent être cités, comme celui de Facebook, ou les réseaux de transport.

Dans ce document, on s'intéresse aux graphes multiparties généralisés, qui constituent une généralisation des graphes biparties, dans lesquelles les noeuds sont grouppés dans deux sous-ensembles disjoints, et les noeuds d'un sous-ensemble ne peut avoir une intéraction qu'avec un noeud de l'autre sous-ensemble. Pour ce type de graphes, les Modèles à Blocs Latents (LBM) (Bar-Hen et al., 2020) constitue une méthode d'inférence très efficace. Ceux-ci viennent comme extension aux Modèles à Blocs Stochastiques (SBM) (Bar-Hen et al., 2020), proposés pour des graphes ne présentant pas de structures biparties ou multiparties.

Les graphes multiparties généralisés sont constituées de noeuds appartenant à plus de deux groupes, appelés groupes fonctionnels. Les interactions peuvent avoir lieu entre les groupes aussi bien qu'au sein de chaque groupe. Ces graphes sont très présents en ethnobiologie (Bar-Hen et al., 2020) où l'une des problématiques centrales est de comprendre la biodiversité des espèces, en rapport avec les interactions qui ont lieu entre les individus.

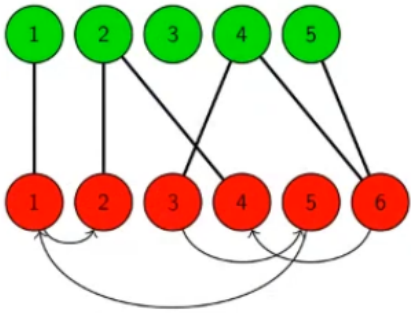

Figure 1. Illustration d'un graphe multipartie généralisé

La procédure d'inférence pour les Modèles à Blocs Multiparties (MBM) repose sur l'introduction de variables latentes et l'utilisation de la version variationnelle de l'algorithme Expectation-Maximization. Ceci aboutit de manière naturelle à la détection de communautés au sein du graphe.

Dans ce document, une formalisation mathématique du Modèle à Blocs Multiparties est donnée, ainsi qu'une introduction à la méthode de sélection de modèle, qui est une version

modifiée de la Vraisemblance Intégrée de Classification (ICL), utilisée pour les SBM. La section 1 est consacrée à l'introduction du modèle et de la procédure d'inférence et on s'intéressera dans la section 2 à la sélection de modèle. Des illustrations numériques sont proposées dans la dernière section.

1. Modèle à blocs multiparties

1.1. Notations

Un graphe multipartie est composé de Q groupes fonctionnels, chaque groupe pouvant être un graphe simple ou bipartie. Soit n_q le nombre de noeuds dans le $q^{ème}$ groupe fonctionnel ($q = 1, ..., Q$). Les paires de groupes fonctionnels sont indexées par $(q, q') \in \{1, ..., Q\}^2$ et on note Σ la liste des paires de groupes pour lesquelles une interaction est observée.

Pour tout $(q, q') \in \Sigma$, on note $X^{qq'}$ la matrice d'interaction, l'entrée $X^{qq'}_{ii'}$ indique l'existence (ou pas) d'une arête entre le noeud i du groupe q et le noeud i' du groupe q'. Les entrées peuvent êtres binaires ($X^{qq'}_{ii'} \in \{0, 1\}$) ou valuées ($X^{qq'}_{ii'} \in \mathbb{R}$ ou $\mathbb{N}$).

La matrice $X^{qq'}$ peut être symétrique, indiquant que le graphe n'est pas orienté, ou non-symétrique dans le cas d'un graphe orienté. $\mathbf{X} = \left(X^{qq'}\right)_{(q,q')\in\Sigma}$ encode l'ensemble du graphe multipartie généralisé. Pour tout $(q, q') \in \Sigma$, on note $\mathcal{S}^{qq'}$ la liste de toutes les interactions possibles entre les groupes fonctionnels q et q'.

1.2. Équations du Modèle à Blocs Multiparties

Le modèle à blocs stochastiques classique est un modèle génératif dans lequel sont introduites n variables latentes Z_i, $i \in \{1, ..., n\}$. Les variables aléatoires Z_i sont supposées indépendantes telles que $\forall i \in \{1, ..., n\}, \forall k \in \{1, ..., K\}$:

$$P(Z_i = k) = \pi_k$$

avec $\pi_k \in [0, 1] \forall k \in \{1, ..., K\}$ et $\sum_{k=1}^{K} \pi_k = 1$.

Les noeuds se connectent alors de la manière suivante:

$$X_{ii'}|Z_i = k, Z_{i'} = k' \sim \mathcal{F}(\alpha_{kk'})$$

où $\mathcal{F}$ est une distribution qui dépend de la nature du graphe, par exemple une Bernoulli dans le cas binaire, Poisson dans le cas de vote ou une Gaussienne dans le cas continu. $\alpha_{kk'}$ représente le vecteur de paramètres de la distribution.

Le modèle à blocs multiparties est la généralisation du SBM dans le cas d'un graphe multipartie. Les équations sont alors modifiées pour prendre en compte l'existence des Q groupes fonctionnels. Nous supposons que chaque groupe fonctionnel est divisé en K_q blocs. Pour tout $q \in \{1, ..., Q\}$ et $i \in \{1, ..., n_q\}$, soit Z^q_i la variable aléatoire latente telle que $Z^q_i = k$ si l'individu i du groupe q appartient au cluster k. Comme dans le cas du SBM, les variables aléatoires Z^q_i sont supposées indépendantes et telles que: $\forall q \in \{1, ..., Q\}, \forall k \in \{1, ..., K_q\}, \forall i \in \{1, ..., n_q\}$:

$$P(Z^q_i = k) = \pi^q_k \tag{1}$$

avec $\pi_k^q \in [0,1]\ \forall k \in \{1,...,K_q\}$ et $\sum_{k=1}^{K} \pi_k^q = 1$. On utilisera par la suite les notations $\mathbf{Z}^q = (Z_{ii'}^q)_{i\in\{1,...,n_q\}}$, $\mathbf{Z} = (\mathbf{Z}^q)_{q\in\{1,...,Q\}}$ et $\pi = (\pi_k^q)_{q\in\{1,...,Q\},\ k\in\{1,...,K_q\}}$.

Ainsi, il y a interaction entre les noeuds si

$$X_{ii'}^{qq'} | \{Z_i^q = k, Z_{i'}^{q'} = k'\} \sim \mathcal{F}(\alpha_{kk'}^{qq'}) \tag{2}$$

Les équations (1) et (2) définissent le modèle à blocs multiparties.

Remarques. *Pour $\Sigma = \{(1,1)\}$ càd $q = 1,\ q' = 1$, on retrouve le modèle à blocs stochastiques classique. De plus, la même structure de dépendance est observée des le modèle à blocs multiparties. En effet, conditionnellement à $\mathbb{Z}$, les entrées $X_{ii'}^{qq'}$ sont indépendantes, mais comme les variables (Z_i^q) sont latentes, la marginalisation engendre de l'indépendance entre les entrées $X_{ii'}^{qq'}$, mais aussi entre les matrices d'interactions $(X^{qq'})_{(q,q')\in\Sigma}$. Par conséquence, les variables latentes $(Z_i^q)_{i\in\{1,...,n_q\},q\in\{1,...,Q\}}$ sont dépendantes dès qu'elles soient conditionnées par $\mathbf{X}$.*

Pour un vecteur de clustering $\mathbf{K} = [K_1, ..., K_Q]$ donné, les paramètres du Modèle a Blocs Multiparties (MBM) sont alors $\theta_{\mathbf{K}} = (\vec{\alpha}, \vec{\pi})$.
Avec $\vec{\alpha} = \left(\alpha_{kk'}^{qq'}\right)_{k\in\{1,...,K_q\},\ k'\in\left\{1,...,K_{q'}\right\},\ (q,q')\in\Sigma}$ et $\vec{\pi} = (\pi_k^q)_{k\in\{1,...,K_q\},\ q\in\{1,...,Q\}}$.

Les tâches suivantes sont alors à réaliser:

1. Pour un $\mathbf{K}$ donné, estimer $\theta_{\mathbf{K}}$
2. Estimer le bon $\mathbf{K}$

La section suivante est consacrée au point 1.

1.3. Vraisemblance du modèle

L'idée est de trouver les paramètres $\theta_{\mathbf{K}}$ qui maximisent la vraisemblance. Pour cela, soit $\mathcal{L}_c(\mathbf{X}, \mathbf{Z}; \theta_{\mathbf{K}})$ la vraisemblance des données complètes $(\mathbf{X}, \mathbf{Z})$.

$$\begin{aligned}\mathcal{L}_c(\mathbf{X}, \mathbf{Z}; \theta_{\mathbf{K}}) &= P_{\theta_{\mathbf{K}}}(\mathbf{Z}) \times P_{\theta_{\mathbf{K}}}(\mathbf{X}|\mathbf{Z}) \\ &= \prod_{q=1}^{Q}\prod_{i=1}^{n_q}\prod_{k=1}^{K_q}(\pi_k^q)^{\mathbb{1}_{Z_i^q=k}} \times \prod_{(q,q')\in\Sigma}\prod_{(i,i')\in\mathcal{S}^{qq'}}\prod_{(k,k')\in\mathcal{A}^{qq'}}(f_{qq'}^d(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'}))^{\mathbb{1}_{(Z_i^q=k, Z_{i'}^{q'}=k')}}\end{aligned}$$

où $f_{qq'}^d$ est la densité de la loi $\mathcal{F}_{qq'}$, si $q \neq q'$ $\mathcal{A}^{qq'} = \{1,...,K_q\} \times \{1,...,K_{q'}\}$ et si $q' = q$ $\mathcal{A}^{qq} = \{1,...,K_q\}^2$ dans le cas dirigé et $\mathcal{A}^{qq} = \{(k,k') \in \{1,...,K_q\} \mid k \leqslant k'\}$ dans le cas non dirigé . Il est important de remarquer que dans le cas où $Q = 1$ $(\Sigma = \{(1,1)\})$, on retrouve bien l'équation de vraisemblance pour le modèle SBM. La log-vraisemblance s'écrit alors:

$$\begin{aligned}\log \mathcal{L}_c(\mathbf{X}, \mathbf{Z}; \theta_{\mathbf{K}}) = &\sum_{q=1}^{Q}\sum_{i=1}^{n_q}\sum_{k=1}^{K_q}\mathbb{1}_{(Z_i^q=k)}\log(\pi_k^q) \\ &+ \sum_{(q,q')\in\Sigma}\sum_{(i,i')\in\mathcal{S}^{qq'}}\sum_{(k,k')\in\mathcal{A}^{qq'}}1_{(Z_i^q=k,\ Z_{i'}^{q'}=k')}f_{qq'}(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'})\end{aligned} \tag{3}$$

où $f_{qq'}$ est le logarithme de la densité de la loi $\mathcal{F}_{qq'}$.
Les variables $\mathbf{Z}$ étant latentes, on peut marginaliser par rapport à toutes les valeurs possibles de $\mathbf{Z}$, notées $\mathcal{Z} = \{(z_i^q)_{i\in\{1,...,n_q\},q\in\{1,...,Q\}|z_i^q\in\{1,...,K_q\}}\}$, pour obtenir la vraisemblance des données incomplètes, soit $\mathcal{L}(\mathbf{X};\theta_{\mathbf{K}})$. Cette intégration, tout comme le modèle à blocs stochastiques classique, est incalculable (notamment quand K_q et n_q augmentent) ce qui nous pousse à opter pour une approximation variationnelle, détaillée dans la section suivante.

1.4. Approximation variationnelle et algorithme V-EM

Les structures de dépendance sont observées dès que l'on cherche à calculer les probabilités des variables latentes conditionnées par les données $P(\mathbf{Z}|\mathbf{X};\theta_{\mathbb{K}})$. Cette dépendance implique une expression non-explicite du E-step de l'algorithme EM.

Pour contrer ce problème, une approximation variationnelle est effectuée. La probabilité $P(\mathbf{Z}|\mathbf{X};\theta_{\mathbb{K}})$ est approchée par une approximation $\mathcal{R}_{\mathbf{X},\tau}$ tirée de distributions paramétrisées par τ forçant l'indépendance entre les variables Z_i^q (c'est ce qu'on appelle l'approximation de champ moyen):

$$\mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z}) = \prod_{q=1}^{Q}\prod_{i=1}^{n_q}(\tau_{ik}^q)^{\mathbb{1}_{(Z_i^q=k)}} \tag{4}$$

où $\tau_{ik}^q = P_{\mathcal{R}_{\mathbf{X},\tau}}(Z_i^q = k)$ sont les paramètres de la distribution avec $\sum_{k=1}^{K_q}\tau_{ik}^q = 1\ \forall q \in \{1,...,Q\}, \forall i \in \{1,...,n_q\}$. Il s'agit alors de choisir les τ_{ik}^q qui minimisent la divergence de Kullback-Leibler $\mathrm{KL}(\mathcal{R}_{\mathbf{X},\tau}||P(.|\mathbf{X};\theta_{\mathbf{K}})$ entre la distribution réelle et l'approximation.

L'idée est de maximiser une borne inférieure de la log-vraisemblance, ce qui engendre la maximisation de celle-ci. Soit $\mathcal{R}_{\mathbf{X},\tau}$ une famille de distributions de probabilités sur $\mathbf{Z}$. On définit $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau})$ une borne inférieure de la log-vraissemblance des données complètes $(\mathbf{X},\mathbf{Z})$ ainsi

$$\begin{aligned}
\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau}) &= \log\mathcal{L}(\mathbf{X};\theta) - KL\left(\mathcal{R}_{\mathbf{X},\tau}||P(.|\mathbf{X};\theta)\right) && (5)\\
&= \log\mathcal{L}(\mathbf{X};\theta) - \mathbb{E}_{R_{\mathbf{X},\tau}}\log\left(\frac{\mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z})}{P(\mathbf{Z}|\mathbf{X};\theta)}\right) && (6)\\
&= \log\mathcal{L}(\mathbf{X};\theta) - \mathbb{E}_{R_{\mathbf{X},\tau}}[\log\mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z})] + \mathbb{E}_{R_{\mathbf{X},\tau}}[\log P(\mathbf{Z}|\mathbf{X};\theta)] && (7)\\
&= \log\mathcal{L}(\mathbf{X};\theta) - \mathbb{E}_{R_{\mathbf{X},\tau}}[\log\mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z})] + \mathbb{E}_{R_{\mathbf{X},\tau}}[\log P(\mathbf{X},\mathbf{Z};\theta)] && (8)\\
&\quad - \mathbb{E}_{R_{\mathbf{X},\tau}}[\log P(\mathbf{X};\theta)] && (9)\\
&= \log\mathcal{L}(\mathbf{X};\theta) - \mathbb{E}_{R_{\mathbf{X},\tau}}[\log\mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z})] + \mathbb{E}_{R_{\mathbf{X},\tau}}[\log P(\mathbf{X},\mathbf{Z};\theta)] && (10)\\
&\quad - \log P(\mathbf{X};\theta) && (11)\\
&= \mathcal{H}(\mathcal{R}_{\mathbf{X},\tau}) + \mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\log\mathcal{L}_c(\mathbf{X},\mathbf{Z},;\theta)] && (12)\\
&\leq \log\mathcal{L}(\mathbf{X};\theta) && (13)
\end{aligned}$$

Avec $\mathcal{H}(\mathcal{R}_{\mathbf{X},\tau}) := -\mathbb{E}_{R_{\mathbf{X},\tau}}[\log\mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z})]$ l'entropie de la loi $\mathcal{R}_{\mathbf{X},\tau}$ dans l'égalité (14). L'inégalité (15) vient du fait que la divergence de Kullback-Leibler est positive.
L'égalité est obtenue si et seulement si $\mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z}) = P(\mathbf{Z}|\mathbf{X};\theta)$, on retrouve alors l'algorithme EM classique.

L'itération t de l'algorithme VEM comprend deux étapes:

- VE Step

$$\tau^{(t)} = \underset{\tau}{argmin}\ KL\left(\mathcal{R}_{\mathbf{X},\tau}||P(.|\mathbf{X};\theta^{(t-1)}\right) \tag{14}$$
$$= \underset{\tau}{argmax}\ \mathcal{I}_{\theta^{(t-1)}}(\mathcal{R}_{\mathbf{X},\tau}) \tag{15}$$

- M Step

$$\theta^{(t)} = \underset{\theta}{argmax}\ \mathcal{I}_{\theta}(\mathcal{R}_{\mathbf{X},\tau^{(t)}}) \tag{16}$$

L'algorithme VEM génère ainsi une suite $(\tau^{(t)}, \theta^{(t)})_{t\geqslant 1}$ qui permet d'augmenter à chaque étape la borne inférieure $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau})$ de la log-vraissemblance $\log \mathcal{L}(\mathbf{X};\theta)$.

Regardons maintenant l'expression explicite de $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau})$:
En se basant sur l'équation (3) de la log-vraissemblance des données complètes $\log \mathcal{L}_c(\mathbf{X},\mathbf{Z};\theta_{\mathbf{K}})$ et l'expression de l'entropie $\mathcal{H}(\mathcal{R}_{\mathbf{X},\tau}) := -\mathbb{E}_{R_{\mathbf{X},\tau}}[\log \mathcal{R}_{\mathbf{X},\tau}(\mathbf{Z})]$ on obtient :

$$\begin{aligned}
\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau}) &= \mathcal{H}(\mathcal{R}_{\mathbf{X},\tau}) + \mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\log \mathcal{L}_c(\mathbf{X},\mathbf{Z},;\theta)] \\
&= -\mathbb{E}_{R_{\mathbf{X},\tau}}\left[\sum_{q=1}^{Q}\sum_{i=1}^{n_q} \mathbb{1}_{Z_i^q=k}\log\left(\tau_{ik}^q\right)\right] + \sum_{q=1}^{Q}\sum_{i=1}^{n_q}\sum_{k=1}^{K_q}\mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k)}\log\left(\pi_k^q\right)] \\
&\quad + \sum_{(q,q')\in\Sigma}\sum_{(i,i')\in\mathcal{S}^{qq'}}\sum_{(k,k')\in\mathcal{A}^{qq'}} \mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_{i'}^{q'}=k')} f_{qq'}(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})] \\
&= -\sum_{q=1}^{Q}\sum_{i=1}^{n_q}\log\left(\tau_{ik}^q\right)\mathbb{E}_{R_{\mathbf{X},\tau}}[\mathbb{1}_{Z_i^q=k}] + \sum_{q=1}^{Q}\sum_{i=1}^{n_q}\sum_{k=1}^{K_q}\mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k)}]\log\left(\pi_k^q\right) \\
&\quad + \sum_{(q,q')\in\Sigma}\sum_{(i,i')\in\mathcal{S}^{qq'}}\sum_{(k,k')\in\mathcal{A}^{qq'}} \mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_{i'}^{q'}=k')}] f_{qq'}(X_{ii'}^{qq'},\alpha_{kk'}^{qq'}) \\
&= -\sum_{q=1}^{Q}\sum_{i=1}^{n_q}\log\left(\tau_{ik}^q\right)\tau_{ik}^q + \sum_{q=1}^{Q}\sum_{i=1}^{n_q}\sum_{k=1}^{K_q}\tau_{ik}^q\log\left(\pi_k^q\right) \\
&\quad + \sum_{(q,q')\in\Sigma}\sum_{(i,i')\in\mathcal{S}^{qq'}}\sum_{(k,k')\in\mathcal{A}^{qq'}} \mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_{i'}^{q'}=k')}] f_{qq'}(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})
\end{aligned} \tag{17}$$

L'expression du logarithme de la densité $f_{qq'}$ de la loi $\mathcal{F}_{qq'}$ dépend du type de réseau :

- Graphe binaire avec $\mathcal{F}(\alpha_{kk'}^{qq'})$ la loi de Bernouilli de paramètre $\alpha_{kk'}^{qq'}$: la densité est $f_{qq'}^d(X_{ii'}^{qq'},\alpha_{kk'}^{qq'}) = {\alpha_{kk'}^{qq'}}^{X_{ii'}^{qq'}} + (1-\alpha_{kk'}^{qq'})^{(1-X_{ii'}^{qq'})}$ et donc la log-densité est $f_{qq'}(X_{ii'}^{qq'},\alpha_{kk'}^{qq'}) = X_{ii'}^{qq'}\log\left(\alpha_{kk'}^{qq'}\right) + (1-X_{ii'}^{qq'})\log(1-\alpha_{kk'}^{qq'})$

- Graphe Poisson avec $\mathcal{F}(\alpha_{kk'}^{qq'})$ la loi de Poisson de paramètre $\alpha_{kk'}^{qq'}$: la densité est $f_{qq'}^d(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'}) = \frac{\alpha_{kk'}^{qq' X_{ii'}^{qq'}}}{X_{ii'}^{qq'}!} e^{-\alpha_{kk'}^{qq'}}$ et donc la log-densité est $f_{qq'}(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'}) = -\alpha_{kk'}^{qq'} + X_{ii'}^{qq'} \log(\alpha_{kk'}^{qq'}) - \log(X_{ii'}^{qq'}!)$
- Graphe Gaussien avec $\mathcal{F}(\alpha_{kk'}^{qq'})$ la loi Gaussienne de paramètre $\alpha_{kk'}^{qq'} = (\mu_{kk'}^{qq'},\ \sigma_{kk'}^{qq'})$: la log-densité est $\log(f_{qq'}^d(X_{ii'}^{qq'}, \mu_{kk'}^{qq'},\ \sigma_{kk'}^{qq'}))$.

La valeur de $\mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_{i'}^{q'}=k')}]$ à la dernière ligne de l'équation (17) varie selon les différents cas possibles :

1. Si $q' \neq q$ et $(q, q') \in \Sigma$ alors $\mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_{i'}^{q'}=k')}] = P(Z_i^q = k, Z_{i'}^{q'} = k') = P(Z_i^q = k)P(Z_{i'}^{q'} = k') = \tau_{ik}^q \tau_{i'k'}^{q'}$.
On notera $\mathcal{E}_q = \{q' \in \{1, ..., Q\} \mid q' \neq q, (q, q') \in \Sigma\}$ pour la suite des calculs.

2. Si $q' = q$, $(q, q) \in \Sigma$ et $i \neq i'\ \forall i \in \{1, ..., n_q\}$ alors $\mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_{i'}^q=k')}] = P(Z_i^q = k, Z_{i'}^q = k') = P(Z_i^q = k)P(Z_{i'}^q = k') = \tau_{ik}^q \tau_{i'k'}^q$.
On notera V_i^q les noeuds du groupe fonctionnels q à l'exception de i' c'est à dire $V_i^q = \{i' \in \{1, ..., n_q\} \mid i' \neq i\}$ pour la suite des calculs.

3. Si $q' = q$, $(q, q) \in \Sigma$ et $i = i'\ \forall i \in \{1, ..., n_q\}$ alors dans le cas où $k = k'$, $\mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_i^q=k)}] = \mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k)}{}^2] = P(Z_i^q = k) = \tau_{ik}^q$, et dans le cas où $k \neq k'$, $\mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_i^q=k)}] = 0$.
On notera V_i^q les noeuds du groupe fonctionnel q à l'exception de q' c'est à dire $V_i^q = \{i' \in \{1, ..., n_q\} \mid i' \neq i\}$ pour la suite des calculs.

Le deuxième terme de la dernière ligne de l'équation (17) peut donc se réécrire ainsi

$$\sum_{(q,q')\in\Sigma} \sum_{(i,i')\in\mathcal{S}^{qq'}} \sum_{(k,k')\in\mathcal{A}^{qq'}} \mathbb{E}_{\mathcal{R}_{\mathbf{X},\tau}}[\mathbb{1}_{(Z_i^q=k,\ Z_{i'}^{q'}=k')}] f_{qq'}(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'})$$

$$= \sum_{q=1}^{Q} \sum_{q'\in\mathcal{E}_q} \sum_{(i,i')\in\mathcal{S}^{qq'}} \sum_{(k,k')\in\mathcal{A}^{qq'}} \tau_{ik}^q \tau_{i'k'}^{q'} f_{qq'}(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'})$$

$$+\ \mathbb{1}_{(q,q)\in\Sigma} \sum_{i=1}^{n_q} \sum_{i'\in V_i^q} \sum_{(k,k')\in\mathcal{A}^{qq'}} \tau_{ik}^q \tau_{i'k'}^q f_{qq}(X_{ii'}^{qq}, \alpha_{kk'}^{qq})$$

$$+\ \mathbb{1}_{(q,q)\in\Sigma} \sum_{i=1}^{n_q} \sum_{k=1}^{K_q} \tau_{ik}^q f_{qq}(X_{ii}^{qq}, \alpha_{kk}^{qq})$$

En conséquence on obtient l'expression explicite de $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau})$ suivante :

$$\begin{aligned}\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau}) = &-\sum_{q=1}^{Q}\sum_{k=1}^{K_q}\sum_{i=1}^{n_q}\log\left(\tau_{ik}^q\right)\tau_{ik}^q + \sum_{q=1}^{Q}\sum_{k=1}^{K_q}\sum_{i=1}^{n_q}\tau_{ik}^q\log\left(\pi_k^q\right)\\ &+\sum_{q=1}^{Q}\sum_{q'\in\mathcal{E}_q}\sum_{(i,i')\in\mathcal{S}^{qq'}}\sum_{(k,k')\in\mathcal{A}^{qq'}}\tau_{ik}^q\tau_{i'k'}^{q'}f_{qq'}(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})\\ &+\ \mathbb{1}_{(q,q)\in\Sigma}\sum_{i=1}^{n_q}\sum_{i'\in V_i^q}\sum_{(k,k')\in\mathcal{A}^{qq'}}\tau_{ik}^q\tau_{i'k'}^q f_{qq}(X_{ii'}^{qq},\alpha_{kk'}^{qq})\\ &+\ \mathbb{1}_{(q,q)\in\Sigma}\sum_{i=1}^{n_q}\sum_{k=1}^{K_q}\tau_{ik}^q f_{qq}(X_{ii}^{qq},\alpha_{kk}^{qq})\end{aligned} \tag{18}$$

Reste maintenant à trouver les valeurs du paramètre $\tau^{(t)}$ de la loi $\mathcal{R}_{\mathbf{X},\tau}$ et du paramètre $\theta^{(t)}$ du modèle MBM dans l'algorithme VEM.

- (VE) Maximisation de $\mathcal{I}_{\theta^{(t)}}(\mathcal{R}_{\mathbf{X},\tau})$ selon τ ($\theta_{(t)}$ étant fixé) :

  On doit trouver $\tau = (\tau_{ik}^q)_{i,k,q}$ tel que $\forall q \in \{1,...,Q\}, \forall k \in \{1,...,K_q\}, \forall i \in \{1,...,n_q\}$ la dérivée du Lagrangien du problème de maximisation $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau})$ sous contrainte $\sum_{k=1}^{K_q}\tau_{ik}^q = 1\ \forall q \in \{1,...,Q\}, \forall k \in \{1,...,K_q\}$ soit nulle. C'est à dire

$$\frac{\partial}{\partial\tau_{ik}^q}\left[\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X},\tau}) + \sum_{q'=1}^{Q}\sum_{i'=1}^{n_{q'}}\lambda_{i'}^{q'}\left(\sum_{k'=1}^{K_{q'}}\tau_{i'k'}^{q'} - 1\right)\right] = 0$$

  Où $(\lambda_{i'}^{q'})_{q'\in\{1,...,Q\},\ i'\in\left\{1,...,n_{q'}\right\}}$ sont les coefficients du Lagrangien. On a ainsi en dérivant l'expression (18) et la contrainte selon τ_{ik}^q on obtient l'équation suivante :

$$\begin{aligned}0 = &-(1+\log\left(\tau_{ik}^q\right)) + \log\left(\pi_k^q\right)\\ &+\sum_{q'\in\mathcal{E}_q}\sum_{i'\in V_i^q}\sum_{k'\in\left\{1,...,K_{q'}\right\}}\tau_{i'k'}^{q'}f_{qq'}(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})\\ &+\ \mathbb{1}_{(q,q)\in\Sigma}\sum_{i'\in V_i^q}\sum_{k'\in\left\{1,...,K_{q'}\right\}}\tau_{i'k'}^q f_{qq}(X_{ii'}^{qq},\alpha_{kk'}^{qq})\\ &+\ \mathbb{1}_{(q,q)\in\Sigma}\mathbb{1}_{(i,i)\in\mathcal{S}^{qq}}f_{qq}(X_{ii}^{qq},\alpha_{kk}^{qq})\\ &+\lambda_i^q\end{aligned}$$

  Ce problème n'a pas de solution explicite, mais d'après *Daudin et al, 2008* la solution vérifie l'équation de point fixe suivante :

$$\widehat{\tau}_{ik}^q \propto \pi_k^q\prod_{i'\neq i}\prod_{q'}\prod_{k'}\mathcal{F}(X_{ii'}^{qq'}\alpha_{kk'}^{qq'})^{\widehat{\tau}_{i'k'}^{q'}}$$

- (M) Maximisation de $\mathcal{I}_{\theta^{(t)}}(\mathcal{R}_{\mathbf{X},\tau})$ selon θ ($\tau_{(t)}$ étant fixé) :
  On obtient $\forall (q,q') \in \Sigma,\ \ \forall (k,k') \in \{1,...,K_q\} \times \{1,...,K_{q'}\}$:

$$\alpha_{kk'}^{qq'} = \frac{\sum_{ii' \in S_{qq'}} X_{ii'}^{qq'} \tau_{ik}^{q} \tau_{i'k'}^{q'}}{\sum_{ii' \in S_{qq'}} \tau_{ik}^{q} \tau_{i'k'}^{q'}}$$

  et $\forall q \in \{1,...,Q\}, \forall k \in \{1,...,K_q\}$ on a

$$\pi_k^q = \frac{1}{n_q} \sum_{i=1}^{n_q} \tau_{ik}^q$$

Les $\widehat{\tau}_{ik}^q$ définissent un soft clustering des noeuds car ce sont des probabilités à valeurs dans $[0,1]$. On peut en déduire un hard clustering par le maximum à posteriori (MAP) défini par:

$$\widehat{Z}_i^q = \underset{k \in \{1,...,K_q\}}{argmax}\, \widehat{\tau}_{ik}^q$$

2. Sélection de modèle par maximisation de la Vraisemblance Intégrée de Classification (ICL)

Le développement réalisé suppose connu le vecteur contenant le nombre de clusters au sein des différents groupes fonctionnels. Dans la majorité des applications, le nombre de clusters $\mathbf{K} = (K_1,...,K_Q)$ est inconnu et doit être estimé. Le critère ICL proposé pour les SBM peut s'étendre ici aux graphes multiparties généralisés, en le modifiant de sorte qu'il tienne compte du caractère multipartie du graphe.

La stratégie sélectionne un MBM à $\mathbf{K}$ clusters $\mathcal{M}_{\mathbf{K}}$. Le critère est le suivant :

$$\text{ICL}(\mathcal{M}_{\mathbf{K}}) = \log \mathcal{L}(\mathbf{X}, \hat{\mathbf{Z}}; \hat{\theta}_{\mathbf{K}}) - \text{pen}(\mathcal{M}_{\mathbf{K}}) \tag{19}$$

Le terme de pénalité contient de l'information et tient compte de l'existence des groupes fonctionnels, il s'écrit ainsi :

$$\text{pen}(\mathcal{M}_{\mathbf{K}}) = \frac{1}{2} \left\{ \sum_{q=1}^{Q} (K_q - 1) \log(n_q) + \left(\sum_{(q,q') \in \Sigma} d_{qq'} |\mathcal{A}^{qq'}| \right) \log \left(\sum_{(q,q') \in \Sigma} d_{qq'} |\mathcal{S}^{qq'}| \right) \right\} \tag{20}$$

Le terme de gauche fait intervenir le nombre de noeuds ainsi que le nombre de clusters pour chaque groupe fonctionnel et contient alors de l'information sur la distribution des clusters. Le terme de droite informe sur les interactions entre groupes fonctionnels. En effet, $|\mathcal{S}^{qq'}|$ est le cardinal de l'ensemble contenant toutes les interactions possibles au sein d'un groupe.

Le meilleur modèle est celui qui maximise le critère (19). Le calcul de ce ritère pour tous le modèles possibles résulte en un calcul intractable. (Bar-Hen et al., 2020) proposent un algorithme pratique pour la sélection de modèle qui, à partir d'un modèle $\mathcal{M}^{(m)}$ donné, considère deux modèles $\mathcal{M}_+^{(m+1)q}$ et $\mathcal{M}_-^{(m+1)q}$ où , pour chaque groupe fonctionnel, la différence est de 1 cluster de plus pour $\mathcal{M}_+^{(m+1)q}$ et 1 cluster de moins pour $\mathcal{M}_-^{(m+1)q}$. Le critère (19) est ensuite calculé pour les deux nouveaux modèles. A chaque itération, l'algorithme sélectionne le modèle qui maximise (19) parmi $\{\mathcal{M}^{(m)}\} \cup \cup_{q \in \{1,...,Q\}} \left\{ \{\mathcal{M}_+^{(m+1)q}\} \cup \{\mathcal{M}_-^{(m+1)q}\} \right\}$. Voir (Bar-Hen et al., 2020) pour l'algorithme exact.

3. Illustrations numériques

3.1. Présentation des données et des paramètres du modèles

Le jeu de données étudié est un réseau écologique représentant les relations mutuelles entre quatre groupes fonctionnels ($Q = 4$) : les plantes ($q = 1$), les pollinisateurs ($q = 2$), les fourmis ($q = 3$) et les oiseaux frugivores ($q = 4$).
Il y a $n_1 = 141$ espèces de plantes, $n_2 = 173$ espèces de pollinisateurs, $n_3 = 30$ espèces de fourmis, $n_4 = 46$ espèces d'oiseaux.

On observe au total 753 interactions dont 55% sont des interactions plantes-pollinisateurs, 28% sont des interactions plantes-fourmis et 17% sont des interactions plantes-oiseaux. Donc le graphe représentant ce réseaux est un graphe multiparti classique où les seules interactions sont entre les plantes et les trois autres groupes fonctionnels. Par conséquent $\Sigma = \{(1,2),(1,3),(1,4)\}$.

L'ensemble des matrices d'interactions est $\mathbf{X} = \{X^{12}, X^{13}, X^{14}\}$ avec $\forall q' \in \{2,3,4\}$, $X_{ii'}^{1q'} = 1$ si la $i^{ème}$ espèce de plante a été observée au moins une fois dans une interaction mutuelle avec la $i'^{ème}$ espèce animal du groupe fonctionnel q' durant la période d'observation, et 0 sinon. Les interactions sont donc binaires et non dirigées.

Nos données ont été stockées sous forme de liste dans MPEcoNetwork. Cette liste a trois éléments qui sont, dans l'ordre, les matrices X^{13}, X^{14} et X^{12} représentant les interactions entre les plantes et les trois autres groupes (fourmis, oiseaux et pollinisateurs).

```
[1] "nombre de matrices d'interactions: 3"
[1] "------------------------------------------------------------"
[1] "noms des matrices d'interaction: "
[1] "Inc_plant_ant"    "Inc_plant_bird"   "Inc_plant_flovis"
[1] "------------------------------------------------------------"
[1] "Début de la première matrice d'intéraction plante-fourmis:"
                        Camponotus_planatus Camponotus_mucronatus Paratrechina_longicornis_
Acacia_cornigera                          0                     0                        0
Acacia_macracantha                        1                     1                        0
Achatocarpus_nigricans                    0                     0                        0
Agave_angustifolia                        0                     0                        0
Amphilophium_paniculatum                  1                     0                        0
```

3.2. Estimation du nombre de clusters et des paramètres du modèle MBM

On estime le nombre de clusters à l'aide de la procédure décrite dans la section 2, avec une liste initale de clusters $K^{(0)} = [1,1,1,1]$ et un nombre maximal de dix clusters pour chaque groupe fonctionnel. Pour cela on utilise la fonction MULTIPARTITEBM.

L'algorithme s'arrête après avoir exploré neuf modèles. Le meilleur modèle, selon le critère ICL, sélectionne la liste de clusters $\mathbf{K} = [7,2,2,1]$. C'est à dire 7 clusters de plantes, 2 clusters de pollinisateurs, 2 clusters de fourmis et 1 cluster d'oiseaux.

La commande FITTEDMODEL de la fonction retourne, pour chaque modèle exploré, une liste de différents paramètres estimés (notamment la liste du nombre de clusters par groupe fonctionnel ainsi que les paramètres $\vec{\alpha}$ et $\vec{\pi}$ du modèle). Les modèles sont ordonnés par ICL décroissant. Le premier modèle (celui ayant l'ICL le plus élevé) fourni les paramètres suivants :

```
[1] "------------Nb of entities in each functional group--------------"
plants flovis   ants  birds
   141    173     30     46
[1] "------------Probability distributions on each network--------------"
[1] "bernoulli" "bernoulli" "bernoulli"
[1] "-----------------------------------------------------------------"
[1] " ------ Searching the numbers of blocks starting from [ 1 1 1 1 ] blocks"
[1] "ICL : -3582.35 . Nb of blocks: [ 1 1 1 1 ]"
  |=============================================================================| 100%, Elapsed 00:00
[1] "ICL : -3358.85 . Nb of blocks: [ 2 1 1 1 ]"
  |=============================================================================| 100%, Elapsed 00:01
[1] "ICL : -3216.14 . Nb of blocks: [ 3 1 1 1 ]"
  |=============================================================================| 100%, Elapsed 00:02
[1] "ICL : -3128.29 . Nb of blocks: [ 4 1 1 1 ]"
  |=============================================================================| 100%, Elapsed 00:02
[1] "ICL : -3071.37 . Nb of blocks: [ 5 1 1 1 ]"
  |=============================================================================| 100%, Elapsed 00:03
[1] "ICL : -3017.65 . Nb of blocks: [ 5 2 1 1 ]"
  |=============================================================================| 100%, Elapsed 00:07
[1] "ICL : -2981.16 . Nb of blocks: [ 5 2 2 1 ]"
  |=============================================================================| 100%, Elapsed 00:10
[1] "ICL : -2965.54 . Nb of blocks: [ 6 2 2 1 ]"
  |=============================================================================| 100%, Elapsed 00:10
[1] "ICL : -2961.9 . Nb of blocks: [ 7 2 2 1 ]"
  |=============================================================================| 100%, Elapsed 00:11
[1] "Best model------ ICL : -2961.9 . Nb of clusters: [ 7 2 2 1 ] for [ plants , flovis , ants , birds ]
respectively"
```

Le clustering réalisé par l'algorithme VEM nous donnes les résultats suivants.

Pour visualiser le réseau on peut afficher la vue mésoscopique. La taille des nœuds est proportionnelle à la taille des clusters et la largeur des arrêtes est proportionnelle à la probabilité de connexion entre ou au sein des clusters. Les arrêtes correspondantes à des probabilités de connexion inférieures à 0,01 ne sont pas tracées.

On peut remarquer que les espèces de plantes des clusters 7 et 2 interagissent uniquement avec les fourmis. Les espèces de plantes des clusters 6 et 3 interagissent exclusivement avec les oiseaux (la différence entre les deux clusters étant la force d'interaction). La différence entre les deux clusters de pollinisateurs tient uniquement à l'existence du cluster 1 des plantes, avec lequel est connecté uniquement le cluster 1 des pollinisateurs.

3.3. Comparaison avec le modèle LBM

On peut se demander si le clustering fourni par le modèle MBM est le même que celui obtenu en analysant chaque réseau biparti séparément avec le modèle LBM. Pour comparer les deux clustering on analyse d'une part le nombre de clusters sélectionnés pour chaque groupe fonctionnel et d'autre part l'Adjusted Rand Index (ARI). L'ARI permet de comparer deux clusters: il est proche de 0 si les deux clusters sont différents et égal à 1 si les deux clusters sont identiques (à une permutation près des classes).

Le modèle LBM trouve 3 clusters de plantes, 3 clusters de pollinisateurs, 2 clusters de fourmis et 1 cluster d'oiseaux. Le clustering des fourmis et des oiseaux est le même pour les deux modèles étant donné que l'ARI est égal à 1. Le clustering des pollinisateurs obtenu par analyse bipartie est légèrement différent (3 clusters contre 2 clusters pour l'analyse jointe des réseaux). Cependant l'ARI est très proche de 1 ce qui signifie que le cluster supplémentaire contient peu d'espèce de pollinisateurs.

Pour les plantes, les trois clusterings correspondants aux trois réseaux bipartis sont très différents du clustering obtenu avec le modèle MBM, les ARI étant respectivement égaux à 0.118, 0.415 et 0.163.

3.4. Efficacité de la procédure d'inférence

References

Avner Bar Hen. Pierre Barbillon, Sophie Donnet (2020): *Block Models for Generalized Multipartite Networks: Applications in Ecology and Ethnobiology*. , Statistical Modelling doi:10.1177/1471082X20963254.